Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality

Thomas Pickard
University of Leeds
Leeds LS2 9JT, UK
mat2tmrp@leeds.ac.uk , tom@tompickard.co.uk

Abstract

This paper explores the use of word2vec and GloVe embeddings for unsupervised measurement of the semantic compositionality of MWE candidates. Through comparison with several human-annotated reference sets, we find word2vec to be substantively superior to GloVe for this task. We also find Simple English Wikipedia to be a poor-quality resource for compositionality assessment, but demonstrate that a sample of 10% of sentences in the English Wikipedia can provide a conveniently tractable corpus with only moderate reduction in the quality of outputs.

1 Introduction

Multiword expressions (MWEs) are word combinations exhibiting one or more idiosyncrasies—lexical, syntactic, semantic, pragmatic or statistical (Sag et al., 2002). This paper is concerned specifically with **semantic compositionality**: the extent to which the meaning of an MWE can be understood from those of its component words. The semantics of compositional expressions such as *picnic basket* are clear to anyone familiar with the constituents *picnic* and *basket*, but a non-compositional phrase like *iron curtain* is opaque without further context.

Word embedding models such as GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013a) and doc2vec (Le and Mikolov, 2014) are widely used in the Natural Language Processing (NLP) sphere, and are capable of capturing syntactic and semantic relationships between words through their representations in multi-dimensional vector space (Mikolov et al., 2013c). These models therefore offer an opportunity to automatically evaluate the compositionality of an MWE candidate by comparing the embedded representation of the complete expression with those of its component words; we may expect that the vectors of more decomposable phrases will be more similar to those of their constituents. Embedding models themselves also benefit from MWE discovery; by treating multi-word expressions as single units, one may obtain higher-quality representations of simplex words (Mikolov et al., 2013b).

Our main aim in this paper is to evaluate the performance of GloVe models for this purpose, in comparison with word2vec. Given that many state-of-the-art NLP applications have adopted BERT embeddings (Devlin et al., 2019), these were also considered. However, BERT's embeddings differ according to the sentence in which a given word appears. Since our methodology requires comparison between the vector representation of MWE candidates and their constituent words, the use of context-dependent embeddings seems inappropriate.

Section 2 outlines relevant past work in this area, in particular that of Roberts and Egg (2018), whose methodology we adapt and whose results provide us with a valuable point of comparison. Our method and resources are described in section 3, including the human-annotated reference sets used to evaluate our scores. Finally, we discuss our findings in section 4.

2 Past Research

Lin (1999) employs a substitution-based method to detect non-compositionality. However, while non-compositional phrases also exhibit **institutionalisation** (resistance to substitution of synonyms), the re-

This work is licensed under a Creative Commons Attribution 4.0 International Licence.
Licence details: http://creativecommons.org/licenses/by/4.0/.

Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 95–100
Barcelona, Spain (Online), December 13, 2020.

verse implication does not hold: institutionalised phrases are not inherently non-compositional (Farahmand et al., 2015). Approaches based on substitution therefore seem better suited to discovery of institutionalised MWEs than to semantically non-compositional ones.

Schone and Jurafsky (2001) and Baldwin et al. (2003) adopt Latent Semantic Analysis (LSA) models based on co-occurrence with 1,000 frequent content words, but more promising results have been obtained through the application of predictive vector embeddings. In particular, the work of Salehi et al. (2015) demonstrated that word embeddings were superior to count-based distribution models when measuring the compositionality of MWEs. Interestingly, they did not find any benefit to using a more complex multi-sense skip-gram (MSSG) model to allow for polysemy of words and expressions . However, their approach was driven by (small) pre-existing lists of MWEs prouced by human annotators.

More recently, Roberts and Egg (2018) generated a large list (over 900k entries) of multi-word phrases, which they extracted from English Wikipedia and automatically scored for compositionality using an approach inspired by Salehi et al. (2015). Our methodology (described in section 3.3) is based on theirs, with alterations to the source corpora and reference sets as well as to the embedding models used.

3 Resources and Methodology

3.1 Corpora

Two training corpora were used, both derived from Wikipedia extracts. In both cases, the XML dumps were processed with a modified corpus reader from the `gensim` Python package (Řehůřek and Sojka, 2010), dividing content articles into sentences and tokens with `punkt` (Kiss and Strunk, 2006) and applying cleansing steps to remove much of the Wiki formatting markup. Note that no case normalisation or lemmatisation was applied.

SIMP20 Complete Simple English Wikipedia content from 2020-06-01. 31,796,513 tokens.

EN20_10P 10% sample of sentences from the 2020-05-20 English Wikipedia. 305,657,697 tokens.

3.2 Reference Sets

Five 'gold standard' lists of MWEs accompanied by compositionality rankings provided by human annotators were employed, providing reference points for intrinsic evaluation of our results. The same reference sets were used by Roberts and Egg (2018), and we also adopt their abbreviated names.

F_ENC (Farahmand et al., 2015). 1,042 nominal compounds (e.g. *greenhouse gas*, *machine language*), with four binary compositionality judgements made by fluent speakers with backgrounds in linguistics. Summing across the judgements produces a four-point scale.

R_ENC (Reddy et al., 2011). 90 noun compounds (e.g. *ivory tower*, *graduate student*), with mean compositionality scores derived from judgements (on a scale from 0 to 5) made by participants recruited through Amazon Mechanical Turk.

MC_VPC (McCarthy et al., 2003). 116 verb-particle pairs (e.g. *space out*, *lie down*), with judgements on a scale from 0-10 made by three judges. The mean of these scores is used, discounting any "don't know" responses. NB: Roberts and Egg (2018) report 117 instances in this dataset, likely due to the presence of a duplicate record which we have removed.

D_ADJN (Biemann and Giesbrecht, 2011). 135 adjective-noun compounds (*blue chip*, *smart card*), taken from the training and test data for the DiSCo 2011 Shared Task. Judgements were made by workers on Amazon Mechanical Turk, averaged and supplied in the range (0,100).

NB: Roberts and Egg (2018) report only 68 instances here. The reason for this is unclear; it may be that additional data were made available by the conference organisers since their work was undertaken. The coverage and correlation measured between their output and this dataset is very similar to that reported in their original paper[1]; we have no reason to believe that this discrepancy has had any negative impact on our findings.

[1] Roberts and Egg (2018) report $\rho = 0.525$, $r = 0.581$ with 64/68 MWEs matching. We obtain, using their published data and matching 118/135 MWEs, $\rho = 0.528$, $r = 0.605$.

MC_VN (McCarthy et al., 2007). 638 verb-object pairs (e.g. *take root*), taken from the list of Venkatapathy and Joshi (2005) and annotated by two judges on a scale from 1 to 6. These two scores are averaged. As Roberts and Egg (2018) point out, many of the pairs are discontiguous (*catch eye*); since our methodology examines only contiguous n-grams, the overlap with this set is restricted.

We also import the automatically-scored list produced by Roberts and Egg (2018), filtering out items which meet the authors' exclusion criteria. This leaves 917,647 items, which we denote by **RE_WIKI15** (since it was derived from the full April 2015 text of English Wikipedia, ca. 2.8 billion words).

3.3 Methodology

We collate corpus frequency counts for contiguous n-grams ($n \leq 3$) and identify MWE candidates by computing the Poisson association measure of Quasthoff and Wolff (2002), adjusting where appropriate to balance it for trigrams. A minimum frequency of 20 occurrences is applied. From the **SIMP20** corpus, we retain the 150,000 most strongly-associated candidate n-grams. For **EN20_10P**, we keep 500,000 items.

In order to enable retokenisation of MWE candidates in the corpora, the n-grams are sorted into distinct batches such that no overlaps are present: the first k words of any n-gram must not be the same as the last k words of any other n-gram in the same batch. A limit of 15 batches is set for **SIMP20** and 10 batches for **EN20_10P**. n-grams consisting entirely of stopwords (the 50 most frequent individual tokens in the corpus) and those which cannot be assigned to a batch are excluded. A total of 148,868 candidates from **SIMP20** and 469,587 from **EN20_10P** were evaluated for compositionality.

For each batch, we replace all instances of the candidate n-grams with a single token and construct `word2vec` (Mikolov et al., 2013a) and `GloVe` (Pennington et al., 2014) word embedding vectors for every simplex word exceeding the minimum frequency of 20, and for all MWE candidates in the batch.

The `word2vec` parameters were those found to be effective by Baroni et al. (2014)[2].

`GloVe` co-occurrence statistics were constructed using a symmetrical window of size 10 without crossing sentence boundaries, and weighted inversely by distance. To maintain tractability, the size of the co-occurrence matrices were restricted by limiting the vocabulary used to the most frequent N simplex words, plus the batch MWE candidates. N was taken to be 300,000, yielding a maximum total vocabulary of size $V = 394,012$ for batch 1 of the EN20_10P corpus. `GloVe` embedding vectors of 300 dimensions were trained with hyperparameters $x_{max} = 100$, $\alpha = 0.75$ and 10 negative samples, as was found to be effective by Pennington et al. (2014). The models were trained for 25 epochs with learning rate 0.05.

Compositionality scores were calculated as the mean cosine similarity between the vector representation of the MWE candidate and each of its component simplex words, ignoring stopwords (we make the assumption that very high-frequency terms are semantically uninformative). The greater the similarity between an MWE and its components, the more semantically transparent the expression.

4 Results

The correlation (Spearman ρ and Pearson's r) between our mean cosine distance measure and human annotations is reported for n-grams appearing on both our list and the reference sets, together with the size of this overlap, in Table 1. We also report the results of Roberts and Egg (2018), using `word2vec` on the full April 2015 English Wikipedia. As there are variances in the **MC_VPC** and **D_ADJN** reference sets, these statistics are recalculated using the authors' published data.

In order to explore the impact of restricting the vocabulary used for training the `GloVe` models, a further experiment was carried out on the **SIMP20** corpus, using an unrestricted vocabulary of 1,014,614 simplex words, together with the MWE candidates assigned to each batch. Table 2 shows the results of this experiment, with the correlations with the reference sets obtained being comparable to those achieved with the `word2vec` embeddings.

[2] Continuous bag-of-words, symmetrical window of size 5. Vectors of length 400 trained over 5 epochs with initial learning rate 0.025, dropping to 0.0001. Negative sampling with 10 samples, subsampling with threshold $t = 10^{-5}$.

Corpus	Model		F_ENC	R_ENC	MC_VPC	D_ADJN	MC_VN
SIMP20	`word2vec`	**Overlap**	179 / 1042	14 / 90	15 / 116	35 / 135	39 / 638
		Spearman ρ	0.169	0.257	0.317	0.316	0.354
		Pearson's r	0.227	0.323	0.398	0.326	0.381
SIMP20	`GloVe`	**Overlap**	183 / 1042	15 / 90	15 / 116	37 / 135	39 / 638
		Spearman ρ	-0.029	-0.061	-0.014	0.234	-0.008
		Pearson's r	-0.135	0.074	0.178	0.231	-0.257
EN20_10P	`word2vec`	**Overlap**	485 / 1042	39 / 90	27 / 116	96 / 135	71 / 638
		Spearman ρ	0.404	0.624	0.536	0.595	0.389
		Pearson's r	0.401	0.632	0.476	0.624	0.366
EN20_10P	`GloVe`	**Overlap**	486 / 1042	39 / 90	27 / 116	96 / 135	71 / 638
		Spearman ρ	-0.043	0.473	-0.122	0.078	-0.188
		Pearson's r	-0.075	0.415	-0.229	0.037	-0.219
WIKI15	`word2vec`	**Overlap**	631 / 1042	61 / 90	47 / 116	118 / 135	132 / 638
		Spearman ρ	0.458	0.615	0.424	0.528	0.392
		Pearson's r	0.473	0.603	0.372	0.605	0.395

Table 1: Correlations between automatically-generated compositionality scores and human-annotated "gold standard" reference lists. The **WIKI15** output is that of Roberts and Egg (2018).

Corpus	Model		F_ENC	R_ENC	MC_VPC	D_ADJN	MC_VN
SIMP20	`GloVe`, full vocab	**Overlap**	183 / 1042	15 / 90	15 / 116	37 / 135	39 / 638
		Spearman ρ	0.200	0.269	0.494	0.101	0.120
		Pearson's r	0.208	0.272	0.492	0.118	0.142

Table 2: `GloVe` model with unrestricted vocabulary on **SIMP20** corpus.

We find substantially lower correlation with the `GloVe`-derived compositionality scores than those obtained using `word2vec`, across both corpora. The `GloVe` model with unrestricted vocabulary appears comparable to `word2vec`, but required greater computational resources to train. Both practical and performance factors lead us to prefer `word2vec` for future work in this area. This aligns with the findings of Baroni et al. (2014) if we regard `GloVe` as an evolution of the 'count-based' vector paradigm, despite its reported success elsewhere (Pennington et al., 2014).

The Simple English Wikipedia corpus produces fewer matches with the reference lists of MWEs as well as weaker correlation with human compositionality judgements; the smaller size of this corpus and the nature of its content make it a poor hunting ground for multi-word expressions. However, our 10% sample of English Wikipedia yielded reasonable results while remaining tractable[3].

Our output lists and code resources are available at `https://github.com/Oddtwang/MWEs` .

Future work includes exploration of context-dependent embeddings such as `doc2vec` (Le and Mikolov, 2014) and BERT (Devlin et al., 2019) for compositionality assessment, particularly for n-grams which may not always form MWEs. Application of the technique to other corpora and languages with suitable MWE resources, e.g. Arabic (Alghamdi and Atwell, 2019) would also be valuable.

Acknowledgements

We are grateful to Dr Eric Atwell of the University of Leeds for his supervision of this work as part of an MSc dissertation project, and for bringing the topic of multi-word expressions to our attention.
Our `GloVe` models used an implementation by Maciej Kula, available from `https://github.com/maciejkula/glove-python`.

[3] Training the `word2vec` models took approximately 2.5 days for 10 batches on the 10% sample of English Wikipedia, using a single Windows desktop PC with an 8-core Intel i7 CPU @ 3.60GHz and 32GB RAM.

References

Ayman Alghamdi and Eric Atwell. 2019. Constructing a corpus-informed list of Arabic formulaic sequences (ArFSs) for language pedagogy and technology. *International Journal of Corpus Linguistics*, 24(2):202–228, August.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment - Volume 18*, MWE '03, pages 89–96, Sapporo, Japan, July. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Chris Biemann and Eugenie Giesbrecht. 2011. Distributional Semantics and Compositionality 2011: Shared Task Description and Results. In *Proceedings of the Workshop on Distributional Semantics and Compositionality*, pages 21–28, Portland, Oregon, USA, June. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A Multiword Expression Data Set: Annotating Non-Compositionality and Conventionalization for English Noun Compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, pages 29–33, Denver, Colorado, June. Association for Computational Linguistics.

Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, 32(4):485–525, November.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196, Beijing, China, June. JMLR.org.

Dekang Lin. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA, June. Association for Computational Linguistics.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan, July. Association for Computational Linguistics.

Diana McCarthy, Sriram Venkatapathy, and Aravind Joshi. 2007. Detecting Compositionality of Verb-Object Combinations using Selectional Preferences. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 369–379, Prague, Czech Republic, June. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*, October.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Uwe Quasthoff and Christian Wolff. 2002. The Poisson Collocation Measure and its Applications. In *Second International Workshop on Computational Approaches to Collocations*, Wien. IEEE.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valetta, Malta, May. ELRA.

Will Roberts and Markus Egg. 2018. A Large Automatically-Acquired All-Words List of Multiword Expressions Scored for Compositionality. In *Proceedings of LREC 2018*. European Language Resources Association (ELRA), May.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 1–15, Berlin, Heidelberg. Springer.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 977–983, Denver, Colorado, May. Association for Computational Linguistics.

Patrick Schone and Daniel Jurafsky. 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Sriram Venkatapathy and Aravind K. Joshi. 2005. Relative Compositionality of Multi-word Expressions: A Study of Verb-Noun (V-N) Collocations. In *Second International Joint Conference on Natural Language Processing: Full Papers*.